# Notitie Beeldbank migratie

Auteurs: Antoinet Nijssen, Maarten Zeinstra, Nico Vriend

Naar aanleiding van de notitie ‘open de sluizen’ is de behoefte ontstaan aan een nadere analyse van de beeldbankconversie die in oktober 2016 heeft plaats gevonden. In totaal werden hierbij 259.902 records geïmporteerd waarvan er momenteel nog 258.831 zijn terug te vinden. De analyse in deze notitie dient er toe om aanknopingspunten te bieden om uiteindelijk te komen tot een afgeronde conversie waarbij het beheer van de beeldcollecties in één collectiebeheersysteem plaats vind. Op basis van de analyse kunnen de acties worden geprioriteerd die noodzakelijk zijn om tot een afronding te komen en welke zaken wellicht in een later stadium worden aangepakt.  
 De notitie begint met een schets van inrichting ontstaan in MF met de conversie gevolgd door overzicht van de belangrijkste aandachtspunten en tenslotte twee mogelijke scenario’s.

## Inrichting MF beeldcollecties

Voorafgaande aan de import in 2016 is er een mapping geweest waarin is bepaald hoe de records vanuit Memorix overgebracht moesten worden naar MF. De documenttypen uit Memorix zijn hierbij gemapt naar de bijpassende archiefeenheidsoorten in MF, onderliggende velden zijn hierbij grotendeels overgenomen. Dat wil zeggen dat er op veld niveau geen daadwerkelijke mapping heeft plaats gevonden maar dat velden als nieuwe flexvelden in MF zijn toegevoegd. Op deze manier zijn aan de hand van de mapping negen (nieuwe) AET’s aangemaakt dan wel heringericht. Hierbij gaat het concreet om:

* Foto (FT)
* Kaart (KRT)
* Prentbriefkaart (PBK)
* Affiche (AFF)
* Tekening (TNG)
* Technische tekening (TTEK)
* Dia (DIA)
* Serie afbeeldingen (SRAFB)
* Schilderij (SHDR)

Waarbij de laatste twee waarschijnlijk pas in een later stadium zijn ontstaan.

Voor een aantal andere gegevens is ervoor gekozen om deze vast te leggen in een contexttoegang of afzonderlijke archiefeenheidsoort.

* Gegevens over vervaardigers zijn vastgelegd in de contexttoegang met mediavervaardigers (con\_mdvdr).
* Straatnaamgegevens zijn vastgelegd in de contexttoegang met straatnamen (con\_strn) waarbij deze toegang wordt gevoed vanuit de AET adres.
* Gegevens over personen op een afbeelding zijn vastgelegd binnen de AET Persoon op afbeelding(PSAFB).
* De tabel trefwoorden in MF is aangevuld met waarden trefwoorden uit Memorix.

## Aandachtspunten

Naast de mapping lijkt er geen verdere uitgebreide analyse van de metadata te hebben plaats gevonden. Op basis van een gecombineerde analyse van de oorspronkelijke data uit Memorix en geïmporteerde data in MF komen de volgende aandachtspunten naar voren.

1. Datavervuiling
2. Structureringsprobleem
3. Ontbrekende bestandskoppeling
4. Ontbrekende handles

## Datavervuiling

Een analyse van het importbestand wijst uit dat de data op meerdere niveaus is vervuild. Concreet gaat het om de volgende twee zaken.

1. Tekstuele vervuiling van metadatavelden.
2. Dubbelen in meerdere varianten.

### Tekstuele vervuiling

Met tekstuele vervuiling wordt bedoeld een verkeerde schrijfwijze, typefouten etc. In MF treffen we deze vervuiling vooral in lijsten en contexttoegangen zoals adressen, mediavervaardigers, trefwoorden en personen op afbeeldingen. Een hoofdrol is hier weggelegd voor de contexttoegangen. Daarnaast is de trefwoordentabel in MF vervuild geraakt doordat trefwoorden vanuit Memorix zonder bewerking of beoordeling zijn toegevoegd.

### Dubbelen in meerdere varianten

Voor de conversie in MF heeft er geen bewerking van de data en metadata plaats gevonden. Dit betekent dat de dubbelen die vóór de conversie reeds aanwezig waren nog steeds bestaan in MF. Na analyse van het originele importbestand kunnen we de volgende varianten in dubbelen onderscheiden: dubbele bestandskoppelingen, dubbel gedigitaliseerd en gekoppeld, identieke dubbelen op basis van metadata.

#### Dubbele bestandskoppelingen

De categorie dubbele bestandskoppelingen bestaat uit bestanden die gekoppeld zijn aan meer dan één record. Uit de analyse van het originele export bestand blijkt dat het hier gaat om 7.101 bestanden en 14.483 records. Nader uitgesplitst is het beeld als volgt.

|  |  |  |
| --- | --- | --- |
| **Bestanden** | **Aantal keer gekoppeld** | **Dubbele records** |
| 6.822 | 2 | 13.644 |
| 277 | 3 | 831 |
| 2 | 4 | 8 |
| 7.101 |  | 14.483 |

Ter illustratie hierbij een voorbeeld waarbij één bestand aan 4 records is gekoppeld. Als je doorklikt op de handles is te zien dat het gekoppelde bestand weliswaar hetzelfde is maar dat de metadata op details afwijkt.

* <https://hdl.handle.net/21.12102/dfaba54a-fb8e-11df-9e4d-523bc2e286e2>
* <https://hdl.handle.net/21.12102/dfab32d6-fb8e-11df-9e4d-523bc2e286e2>
* <https://hdl.handle.net/21.12102/dfab6c92-fb8e-11df-9e4d-523bc2e286e2>
* <https://hdl.handle.net/21.12102/dfad42a6-fb8e-11df-9e4d-523bc2e286e2>

#### Dubbel gedigitaliseerd

Een andere categorie zijn de dubbel gedigitaliseerde of dubbel gekoppelde items binnen één record. Hierbij gaat het om records waarin twee maal een afbeelding voorkomt van hetzelfde fysieke item. Het kan dan gaan om een item dat twee keer is gedigitaliseerd of een bestand dat twee keer is geüpload naar Memorix en vervolgens nogmaals gekoppeld. Van dit type record zijn er 2.724 geïdentificeerd. Hier een voorbeeld van een dergelijk record. <https://hdl.handle.net/21.12102/feb5cdda-fb8e-11df-9e4d-523bc2e286e2>

#### Dubbelen op basis van metadata

Binnen deze categorie vallen dubbelen die op basis van overeenkomsten in de metadata kunnen worden geïdentificeerd. Volledig identieke dubbelen zijn hierin nauwelijks aan te treffen. Wanneer we echter een vergelijking uitvoeren op basis van veelgebruikte velden dan kunnen we in ieder geval concluderen dat er ook binnen de metadata veel dubbelen kunnen worden geïdentificeerd. Beperken we ons alleen tot de velden *inventarisnummer*, *fotonummer* en *collectie* dan zijn er in totaal 20.744 records met 2 of meer keer dezelfde eigenschappen. Breiden we de te analyseren velden uit met *beschrijving* dan tellen we nog 2.320 records die 2 of meer keer dezelfde eigenschappen delen. Bovenstaande toont aan dat een inventarisnummer meerdere keren voorkomt maar een andere beschrijving kent. In deze analyse zijn nog niet de inventarisnummers meegenomen die deels met elkaar overeenkomen omdat er bij het ene inventarisnummer wel een letter voor de omvang is meegenomen en bij de andere niet. Naar alle waarschijnlijkheid zal deze nog meer dubbelen naar voren brengen.

## Structureringsprobleem

Ook doet er zich een structureringsprobleem voor in de metadata. De kern van dit probleem is terug te brengen tot één vraag namelijk; Wat is een beeldbankobject en hoe moet deze worden beschreven? Op deze vraag is niet direct een éénduidig antwoord te geven. Neem bijvoorbeeld een bouwtekening uit een bouwvergunningsdossier. Is de individuele tekening daaruit dan een beeldbankobject? Of is het gehele dossier het eigenlijk te beschrijven object en is de tekening slechts een onderdeel daarvan?   
 Zouden we het beschrijven van een archief als uitgangspunt nemen dan zou een stuk of object worden gedefinieerd als: het geheel van samenhangende gegevens vastgelegd op één of meerdere gegevensdragers. Concreet betekent dit dat een stuk weliswaar uit meerdere delen kan bestaan maar dat het samen één geheel vormt. Dit geheel krijgt een inventarisnummer toegekend. MF is een systeem dat is opgezet om gegevens over archieven vast te leggen en hanteert in principe het bovenstaande uitgangspunt.   
 Bij de beeldcollecties is dit uitgangspunt niet toegepast wat zowel bij de conversie als bij de koppeling van bestanden tot problemen kan leiden. Dit is het beste te illustreren aan de hand van een voorbeeld. <https://hdl.handle.net/21.12102/f857ec8e-fb8e-11df-9e4d-523bc2e286e2> In dit voorbeeld zien we een verzameling tekeningen van de batterijen aan de Karnemelksloot in Naarden. Deze tekeningen tezamen vormen het te beschrijven beeldbankobject. Aan dit beeldbankobject zou normaalgesproken het inventarisnummer toegekend moeten zijn. Zoals we kunnen zien is dit niet het geval maar is er aan elk object afzonderlijk een inventarisnummer toegekend. Om toch de samenhang duidelijk te maken is er in het veld inventarisnummer een reeks inventarisnummers opgenomen die worden onderscheiden door <br> te gebruiken. Deze notatie is niet wenselijk in MF omdat er een verwijzing wordt gemaakt naar meerdere inventarisnummers. Of al deze opgesomde inventarisnummers moeten samen één inventarisnummer gaan vormen. Of ze moeten afzonderlijk worden opgenomen waardoor er dubbele beschrijvingen in MF terecht komen. Dergelijke constructies komen ruim 5322 keer voor.

## Ontbrekende bestandskoppelingen

Bij de conversie van Memorix naar MF is er onvoldoende rekening gehouden met het behouden van de bestandskoppelingen. Deze koppelingen zijn een voorwaarde voor online publicatie. In een later stadium is daarom gekozen voor een collectiegewijze aanpak waarbij de bestanden werden hernoemd op basis van het stramien <*Isilcode>\_<Toegangscode>\_<inventarisnummer>*. De koppeling wordt vervolgens tot stand gebracht door in het codeveld identieke naamgeving toe te passen. Een ideale oplossing om op korte termijn tot een snelle bestandskoppeling te komen. Op de langer termijn zal deze koppelingsmethode problemen met zich meebrengen omdat deze afwijkt van standaardnaamgeving en koppeling.

1. Binnen de beeldcollecties wordt het inventarisnummer nu gelijk gesteld aan één bestand. Consequentie hiervan is dat een volgnummer ontbreekt in de naamgeving waardoor de koppeling van meerdere bestanden problematisch is en ook de juiste volgorde niet goed is te achterhalen.
2. Het opnemen van het stramien <*Isilcode>\_<Toegangscode>\_<inventarisnummer>* in het codeveld maakt standaardkoppeling onmogelijk.
3. Bij koppeling op basis van waarden in het codeveld wordt er geen scantoegang aangemaakt. In plaats daarvan wordt het bestand direct gekoppeld aan het record. Dit betekent dat bepaalde functionaliteiten binnen MF zoals de nadere toegang niet meer voor deze collecties kunnen worden gebruikt.
4. De koppelmethode introduceert een tweede naamgevingsmethode op de NAS. Dit is uit beheeroogpunt op de lange termijn niet wenselijk.

## Handles en identificatie

Binnen de raadpleegomgeving van Memorix worden de handles van handle.net gebruikt als persistent identifier (PID). De aanwezigheid van de handle is een belangrijke voorwaarde voor presentatie en identificatie. Deze PID is samengesteld uit de URL van handle.net gevolgd door de identificatiecode (21.12102) van het Noord-Hollands Archief en tenslotte de UUID uit Memorix. Zoals in het voorbeeld hiernaast is te zien. <https://hdl.handle.net/21.12102/febce610-fb8e-11df-9e4d-523bc2e286e2> De handles vormen daarmee het unieke identificatiekenmerk van de records. De complete handles zijn als zodanig niet terug te vinden in Memorix maar met bovenstaande gegevens eenvoudig te reconstrueren.   
 Binnen Memorix worden UUID’s (Universally unique identifier)gebruikt als uniek identificatiekenmerk van objecten binnen de database. Binnen MF wordt de GUID toegepast als uniek identificatiekenmerk. De GUID is een variant op de UUID. Bij de conversie van Memorix naar MF is niet de UUID van het record opgenomen maar de UUID van het gekoppelde bestand. Deze kunnen niet worden gebruikt als uniek identificatiekenmerk van het record omdat zoals we eerder kunnen hebben zien meerdere bestanden aan één record zijn gekoppeld. Deze UUID’s zijn daarmee niet uniek. Ook bestaan er records waaraan geen bestanden gekoppeld zijn en dus geen UUID meegekregen hebben. In het geval van het laatste is er door MF een nieuwe GUID gegenereerd.   
 Reconstructie van de handles wordt door deze ontstane situatie lastig maar niet onmogelijk. Zoals gezegd is de handle een belangrijk en onmisbaar onderdeel voor de identificatie en presentatie. Presentatie zonder handles is niet wenselijk en daarom is het belangrijk om deze situatie te herstellen. Op basis van de originele export kunnen we constateren dat er 209.860 bestanden een unieke koppeling kennen met één record. Dat betekent dat aan deze bestanden exact 1 record te koppelen is. Dit betekent dat 80,7% van het totaal aantal record herleid kan worden naar de oorspronkelijke UUID uit Memorix.

## Mogelijke scenario’s

Er zijn op basis van bovenstaande analyse twee scenario’s mogelijk om deze problematiek aan te pakken. Aan de hand van de keuze voor een van deze scenario’s kan een actielijst opgesteld worden om de bovenbeschreven problematiek op te lossen.

### Conversie opnieuw

Het volledig uitvoeren van de conversie betekent dat de oude conversie volledig verwijderd zal moeten worden uit MF. Dit brengt als nadeel met zich mee dat reeds gedane arbeid ongedaan wordt gemaakt. Daarnaast zullen sommige onderdelen niet meer verwijderd kunnen worden omdat deze niet kunnen worden geïdentificeerd of te grote complicaties met zich kunnen brengen voor het functioneren van MF. Een volledig nieuwe conversie brengt als voordeel met zich mee dat de belangrijkste identifiers correct kunnen worden meegenomen en dat voorafgaand een goede afstemming kan worden gedaan wat betreft inrichting, bestandskoppeling en koppeling met de standplaatslocatie.

### Conversie behouden

Het behouden van de conversie betekent dat er geen hernieuwde import plaats vindt maar dat er herstelwerkzaamheden binnen MF worden uitgevoerd. De belangrijkste hiervan is het herstellen van de GUID binnen MF en handles binnen de presentatie omgeving. Daarnaast is het tot stand brengen van de bestandskoppeling een voorwaardelijke voorwaarde om tot presentatie te komen. Deze kan op twee manieren tot stand worden gebracht. Via de tot nu toe gehanteerde methode waarop de NAS de juiste bestandstructurering handmatig wordt aangebracht waarna koppeling volgt. Deze methode heeft als nadeel dat het tijdrovend is en bovendien onduidelijk is of de bestanden op de NAS compleet zijn. Of er wordt gekozen voor een alternatieve methode waarbij met een kopie van de bestanden op de DAM een koppeling tot stand kan worden gebracht. Met als voordeel dat er relatief snel gepresenteerd kan worden en herstelwerkzaamheden op de achtergrond kunnen plaats vinden. Daarnaast is het ontdubbelen een belangrijk actiepunt.